We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

11

Quantitative Organelle Proteomics of Protein Distribution in Breast Cancer MCF-7 Cells

Amal T. Qattan and Jasminka Godovac-Zimmermann
Molecular Cell Dynamics, Division of Medicine, University College London, UK

1. Introduction

In his address on the treatment of breast cancer, delivered in 1894 before the Harveian Society of London, W. Watson Chayne said of breast cancer: the "subject cannot be too often brought before the notice of the medical public. First, because the disease is common, at any rate in certain regions, and seems to be becoming more so"(Cheyne 1894). Although a hundred years of extensive research generated 226 946 scientific publications in the period 1886-2011, breast cancer remains the second leading cause of cancer deaths in women today. Breast cancer is the first human tumor for which targeted therapies have been developed. The most successful therapies include tamoxifen and aromatase inhibitors – both estrogen receptor pathway downregulators – and Herceptin, a HER2 antagonist that prolongs disease remission in selected women, but metastatic breast cancer remains largely an incurable disease (Imyanitov and Hanson 2004).

Breast cancer shares all the hallmarks of cancer postulated by Hanahan and Weinberg (Hanahan and Weinberg 2000) that include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. In addition, recent progress has added two further hallmarks such as reprogramming of energy metabolism and evading immune destruction (Hanahan and Weinberg 2011). Increasing recognition of the contribution of the tumor microenvironment to tumorigenesis re-affirms the concept of cancer as a systemic disease with very complex, not yet understood biology.

In 2011 more than 7 million humans around the world will die of cancer and 465 000 women will die from breast cancer alone (Mukherjee 2011).

For humanity, cancer is still the "Emperor of all maladies, master of all terrors."(Mukherjee 2011). For scientists it remains a formidable challenge to understanding the complexity of cellular function.

2. Large-scale proteomics analysis of cancer cells

The completion of human genome sequencing and rapid development of DNA-microarray technology led to extensive investigation of gene expression associated with breast cancer. Large-scale screening of mRNA levels showed that multiple and extensive changes in mRNA levels are commonly seen in breast cancer (Ince and Weinberg 2002, Sebastian and Johnson 2006, Sorlie et al. 2001, van de Vijver et al. 2002). However, eukaryotic cell

proliferation is known to involve complex molecular choreography of mitogens that stimulate cell growth, membrane receptors, their signaling pathways, and downstream effectors of cell division (Hanahan and Weinberg 2000, Sebastian and Johnson 2006). Such studies clearly indicated an urgent need for complementary, highly parallel studies at the protein level. If genes have "legislative power" much of "executive power" is carried out by proteins. Their spatial and temporal distribution within cells is a very complex, but essential, feature of cellular function. The analysis of such distributions is complicated by the facts that a given protein may have multiple subcellular locations, can exist in multiple transcriptional or post-translational isoforms within the same cell and that the different isoforms may have different spatial and temporal distributions as well as different functional roles (Godovac-Zimmermann et al. 2005, Roberts and Smith 2002). Highly parallel methods such as analysis of mRNA abundance can give information on inputs to cellular protein abundance but the mRNA methods do not always correlate well with direct measurements of protein abundance (Gygi et al. 1999), require additional complexity to measure transcriptional isoforms, do not detect post-translational isoforms, and do not give information on spatial location. Conversely, direct measurements of spatial location by methods such as fluorescence microscopy usually do not distinguish isoforms, are mainly semiquantitative, and are difficult to achieve in highly parallel formats.

2.1 Quantitative proteomics of MCF-7 breast cancer subcellular organelles

In recent years, considerable effort has been devoted to determining the identities of proteins included in different subcellular organelles by proteomics (Au et al. 2007, Rogers and Foster 2007, Simpson and Pepperkok 2006, Xu et al. 2009, Yates et al. 2005). The most common approach has been purification of individual organelles followed by exhaustive determination of the protein content. The main disadvantages of this approach are (a) that the degree of purification/contamination of the organelle is difficult to ascertain conclusively for lower abundance proteins, (b) that the protein content may be altered by the purification process and (c) that the approach is not very suitable for dynamic studies of protein subcellular location. In a few cases, (Dunkley et al. 2004, Foster et al. 2006) an alternative approach of partial purification of organelles in a sucrose gradient has been employed, but the assignment of proteins to individual organelles has been based on matching gradient profiles of proteins to the profiles of presumptive marker proteins. Although this is useful for identifying what might be denominated core proteins of an organelle, it is automatically biased against evaluation of proteins in multiple subcellular locations.

The goal of our work (Qattan et al. 2010) was to establish high throughput proteomics methods that are capable of analyzing dynamically at least some of the complexity involved in subcellular protein distribution. The estrogen-dependent MCF-7 malignant breast epithelial cell line was selected due to the wealth of information available in the literature and its relevance to breast cancer (Lacroix and Leclercq 2004, Soule et al. 1973). Proteomics methods based on mass spectrometry are only suitable for indirect measurements of spatial location and we have therefore concentrated on the distribution of proteins between different subcellular organelles. To avoid the need for multiple purification procedures for many different organelles, partial purification based on sucrose gradient centrifugation was used followed by high throughput proteomics analysis of the protein content of different fractions from the sucrose gradient. Figure 1 illustrates the subcellular proteomics

workflow. Following subfractionation of cellular organelles by sucrose gradient centrifugation, the basic functioning of the method was controlled by biochemical assay (Figure 2).

Enzymatic assays and Western blot detection indicated sucrose gradient fractions enriched in cytosol, plasma membrane, endoplasmic reticulum, and mitochondrial proteins, respectively. On the basis of this data obtained, fractions of cytosol, plasma membrane, endoplasmic reticulum and mitochondria from the sucrose gradient fraction were subjected to detailed analysis of protein content by MS methods.

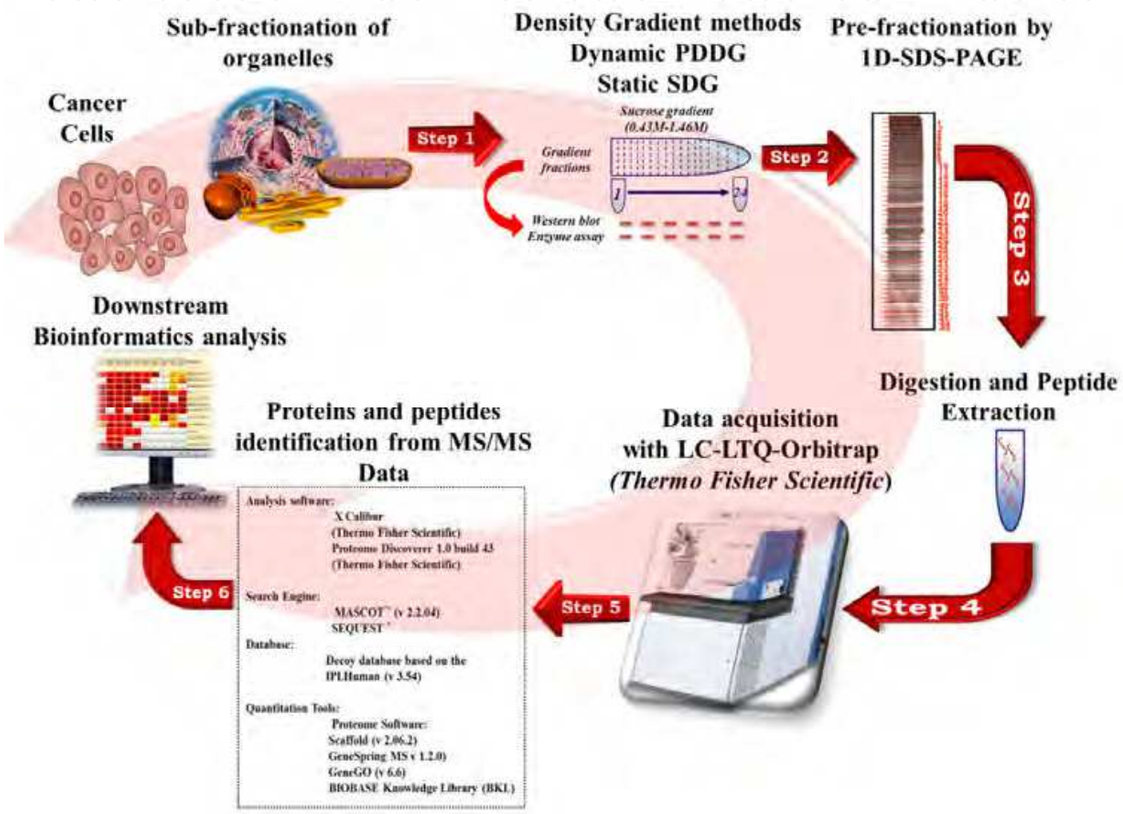


Fig. 1. Subcellular organelle proteomics workflow.

2.1.1 Proteomics data sets show multiple subcellular locations of proteins

Two aspects of the MS analysis are important in the context of the goals of the present work: (1) secure identification of as many proteins as possible in each fraction and (2) accurate measurement of the (relative) amount of any specific protein across the different fractions. We have used direct spectral counts from MS/MS runs for quantitative measurements of the peptides (Usaite et al. 2008). Table 1 shows that that a total of 15 527 different peptides were used to identify 2184 proteins in fractions of cytosol (CT), plasma membrane (PM), endoplasmic reticulum (ER) and mitochondria (MT). The initial set of MS data contained 5514 (protein, fraction, abundance) data points for 2184 proteins: there was an average of 2.5 locations per protein. This *initial data* set contained a substantial number of (protein, fraction, abundance) data points for which in a particular fraction only a single peptide with a small number of spectral counts was observed for some proteins.

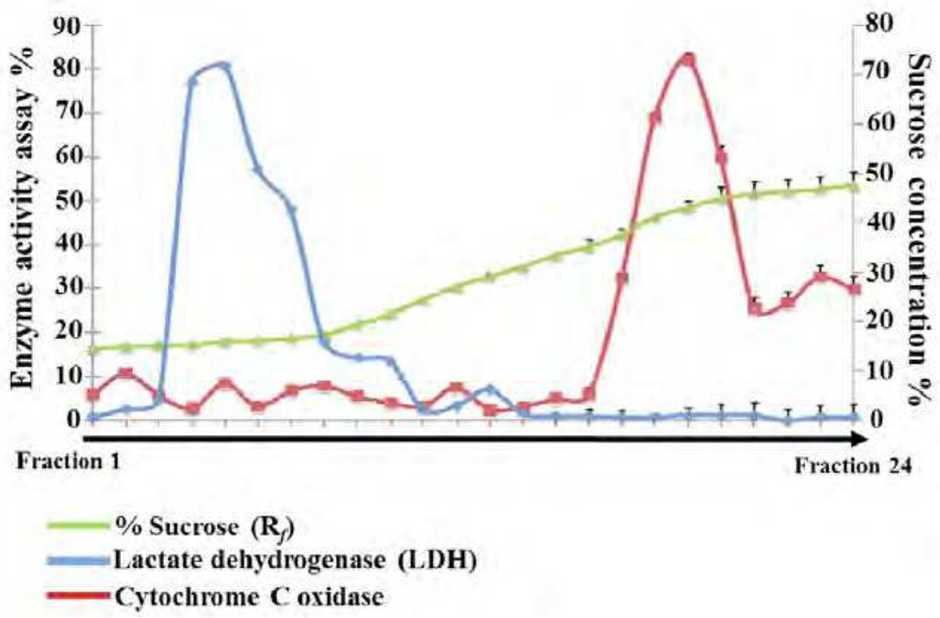


Fig. 2. Subcellular verification of protein markers by enzyme assay and Western blot. A total of 24 fractions was collected from a sucrose-density gradient (0.43-1.46 M) in which fractions 1 and 24 represent the top and the bottom, respectively. Typical density profile of sucrose fractions is calculated by refractive indices. Also shown is the distribution of activity across the gradient for the subcellular enzyme markers LDH and Cytochrome C oxidase.

The assignment of these proteins is less certain for these fractions. Removal of 876 data points for fractions where only a single peptide and 1 or 2 spectral counts were observed gave the *normal data* set in Table 1. 106 (protein, fraction, abundance) data points with only a single peptide in a fraction, but with 3 to 74 spectral counts, were retained to give a total of 4638 data points. The normal data set, which was used for many of the analyses below, corresponds to an average of 2.1 locations per protein. For some of the analyses, we have also removed from the normal data set those (protein, fraction, abundance) data points where less than 4% of the total amount of a given protein was observed in a specific fraction. This *trimmed data* set reduced the number of data points to 4576, that is, an average of 2.1 locations per protein.

In the following we will refer to the three sets of (protein, fraction, abundance) data points used for further analysis as the initial, normal and trimmed data sets (Table 1B), all of which contain a total of 2184 proteins. For individual proteins that were detected in multiple fractions, we will also use the term "primary location" to refer to the (protein, fraction) pair with the highest abundance and the term "secondary location" to refer to other (protein, fraction) pairs with lesser abundances for the same protein.

2.1.2 The observation of the same protein in multiple fractions is not due to "tailing" of the proteins in the sucrose gradient

With the normal data set, many of the proteins were observed in more than one sucrose gradient fraction and hierarchical clustering was used to analyze their distribution over the gradient (Figure 3). This indicated that in many cases the observation of the same protein in multiple fractions was not due to "tailing" of the proteins in the sucrose gradient.

Summary of MS Data				
Distribution Over Sucrose Gradient Fractions				
Fractions				
Total number of	**CT**	**PM**	**ER**	**MT**
Initial Data[a]				
Unique MS Spectra	4393	11435	8628	9654
Unique Peptides	3969	9876	7553	8588
Total Proteins per Fraction	852	1611	1441	1610
Unique Proteins per Fraction[b]	129	116	27	209
Normal Data[c]				
Unique MS Spectra	4233	11233	8341	9427
Unique Peptides	3810	9674	7267	8359
Proteins per Fraction	962	1409	1154	1383
Unique Proteins per Fraction[b]	189	239	69	347
Trimmed Data[d]				
Unique MS Spectra	4092	11223	8320	9375
Unique Peptides	3669	9664	7246	8311
Proteins per Fraction	657	1405	1145	1369
Unique Proteins per Fraction[b]	189	239	69	350

Data Sets		
Data set	**Number of (protein, fraction, abundance) data points**	**Number of proteins**
Initial[a]	5514	2184
Normal[c]	4638	2184
Trimmed[d]	4576	2184

a Includes all (protein, fraction, spectral counts) data points verified by Scaffold.
b Number of proteins found only in one fraction.
c Excludes(protein, fraction, spectral counts) data points where only a single peptide with 1 or 2 spectral counts was observed in a specific fraction.
d After removal from the normal data set of (protein, fraction, abundance) data points for which the proportion of the protein in a specific fraction was less than 4% of the total protein abundance in all four fractions.

Table 1. Summary of MS Data

The data shows numerous examples of bimodal distribution of proteins over two fractions that are not adjacent in the gradient (e.g., cytosol and mitochondria fractions in Figure 3C), as well as examples of proteins with more complicated bimodal distributions over three of

the four fractions (Figure 3B) that are highly unlikely to arise from tailing. A Venn diagram (Figure 4) has been used to summarize the observed distribution of the proteins over the four sucrose gradient fractions as determined by the hierarchical clustering. A notable characteristic for the normal data set is that only 844 of the 2184 proteins (38.6%) were uniquely found in a single fraction. A further 296 proteins (13.6%) were found to be ubiquitously distributed over all fractions. The remaining 1044 proteins (47.8%) were consistent with intermediate distribution over multiple, but not all, subcellular locations. Of these 1044 proteins, 248 (11.4% of total proteins) were distributed over two fractions (e.g., cytosol and mitochondria, Figure 3C) or over three fractions (e.g., cytosol, membrane proteins and mitochondria, Figure 3B) in a "bimodal" manner that is inconsistent with inclusion in a single subcellular organelle and "tailing" over the sucrose gradient.

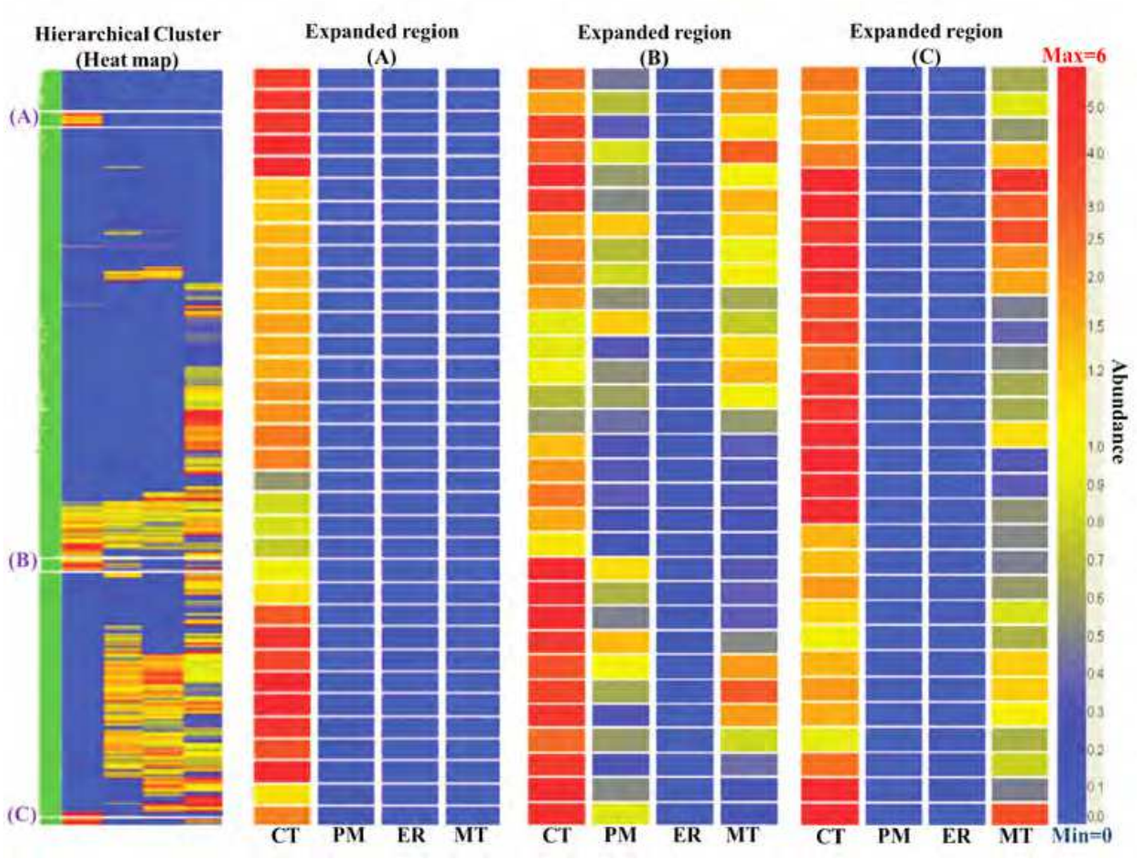


Fig. 3. Hierarchical clustering and heat map across the four fractions. Individual proteins are represented by a single row, each fraction is represented by a single column, and each cell represents the abundance of a single protein in a single fraction. The color scale is for normalized relative abundance from 6.0 (red) to 1.0 (yellow) to 0.0 (blue, not detected). The expansions show typical regions of the heat map corresponding to: (a) proteins observed uniquely in cytosol, (b) "bimodal" proteins (see text) observed in fractions cytosol (CT), plasma membrane (PM), mitochondria (MT), and (c) "bimodal" proteins observed in fractions cytosol and mitochondria.

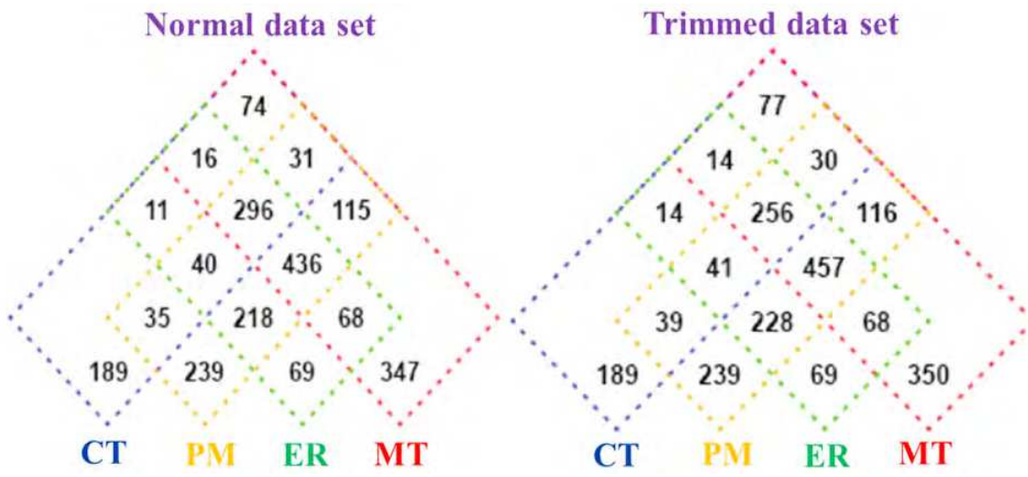


Fig. 4. Four-way Venn diagrams summarizing the distribution of the 2184 proteins over different combinations of the sucrose gradient fractions.

2.1.3 The data represent a good sampling of the distribution over multiple subcellular locations for the observed proteins

Inspection of the distribution of the proteins between primary and secondary locations revealed that they are well dispersed over the regions compatible with a primary location and 1-3 secondary locations (Figure 5). Thus, for example, proteins for which we detected a primary location and a single secondary location must lie on the line from (0.5, 0.5) to (1.0, 0.0) (green plus signs in Figure 5), but are well dispersed along that line. For 2-3 secondary

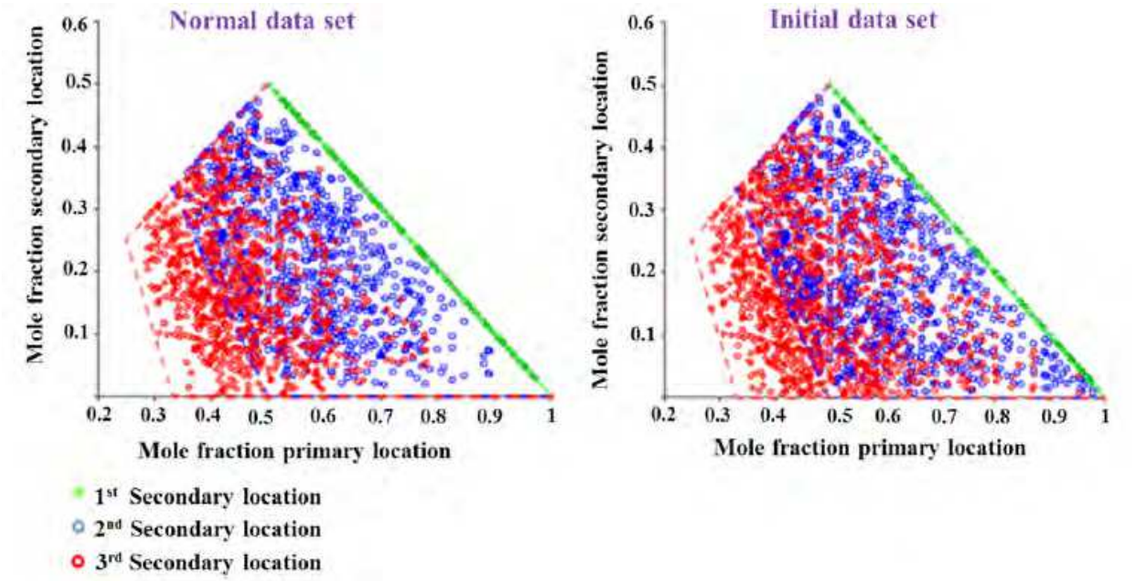


Fig. 5. Distribution of proteins with a primary location and 1 (green), 2 (blue), or 3 (red) secondary locations over compatible areas of a plot of primary mole fractions vs secondary mole fractions. For each protein, the spectral counts observed in a specific gradient fraction were expressed as mole fractions of the total number of spectral counts observed in all four gradient fractions.

locations, the initial data set shows better sampling near the edges of the compatible regions, for example, there are more data points at large values of the primary mole fraction and at very small values of the secondary mole fractions. Many of these data points arise from proteins corresponding to sequencing of only one peptide and only 1-2 spectral counts in a specific fraction. This is a consequence of the sampling properties of spectral counting. The dispersion of the data points in Figure 5 over the compatible areas of the plot is a strong indication that the data represent a good sampling of the distribution over multiple subcellular locations for the observed proteins.

2.1.4 Spurious tailing of proteins in the sucrose gradient does not make any major contributions to the observed multiplicity of locations

A more quantitative evaluation of the possibility of tailing in the gradient was obtained by looking for proteins with high abundance in a given gradient fraction, but with no detectable abundance in the adjacent fractions. For the most abundant proteins, the MS detection method was capable of detecting as little as about 0.2% of the protein in an adjacent fraction. Because the proteins may correspond to different subcellular organelles, tailing between two fractions need not be symmetrical, e.g. tailing from CT to PM may not be the same as tailing from PM to CT. This leads to the six tests for the possibility of tailing shown in Table 2. For all the fractions there are many highly abundant proteins which do not tail into the adjacent fraction (Table 2). The highly abundant proteins also reveal some characteristics which are common in the data set. Some very abundant proteins were found uniquely in a single fraction (e.g., see hepatoma-derived growth factor and Protein S100-A9 in Table 2). Other proteins were detected in only two fractions, but with a bimodal distribution over the fractions (e.g., see sialic acid synthase and pyridoxal kinase in Table 2).
Many proteins were distributed over several fractions, with substantial proportions of the protein present in different fractions (e.g., see ATP-citrate synthase in Table 2). Some proteins were primarily present in a single fraction, but small amounts of the protein were found in other fractions (see e.g. Rho GDP-dissociation inhibitor 1 and nucleophosmin in Table 2). We conclude from the data in Table 2 that spurious tailing of proteins in the sucrose gradient does not make any major contributions to the observed multiplicity of locations.

2.1.5 Annotations of subcellular location

We have used previous subcellular location annotations in the UniProtKB database (in the keyword "subcellular location" field and the ontology "subcellular component" field) and in the Locate Subcellular Location database to compare three aspects of the present work with earlier work: (1) the degree to which the individual sucrose gradient fractions are enriched with proteins corresponding to specific subcellular organelles; (2) the extent to which the multiplicity of subcellular locations observed here is reflected in current annotations of subcellular locations; and, (3) the extent to which there are discrepancies between this work and previous annotations of subcellular locations.
In evaluating these comparisons, it is important to keep in mind that there is not an exact mesh between our experimental strategy and the ontological descriptions of subcellular location used in the databases. The top level of our experimental design matches the levels (extracellular region, plasma membrane, cytoplasm, nucleus) in the GO classification scheme, but the experiment excludes the extracellular region and the nucleus. At a lower level we only tried to obtain an approximate resolution of the cytoplasm as (cytosol,

Test for Overlap of Proteins Between Sucrose Gradient Fractions					
	Normalized protein abundance [b]				
Accession number	CT	PM	ER	MT	Proteins name
					Overlap from CT to PM
	Top [c]	*ND*[c]	*all*[c]	*all*[c]	
4758516	15	-	-	-	**Hepatoma-drived growth factor**
157829557	11.92	-	-	0.77	Carbonic anhydrase2
36038	11.76	-	-	0.49	Rho GDP-dissociation inhibitor 1
4502105	10.97	-	-	0.31	Annexin A4
4506387	9.53	-	0.24	1.47	UV excision repair protein RAD23 homologue B
7023053	7.79	-	-	1.95	**Sialic acid synthase**
4505701	7.37	-	-	1.28	**Pyridoxal Kinase**
					Overlap from PM to CT
	ND[c]	*Top*[c]	*all*[c]	*all*[c]	
24307879	-	7.25	1.58	0.42	Cytoplasmic dynein 1 intermediate chain 2
68533125	-	6.55	1.28	0.72	**ATP-citrate synthase**
34366439	-	6.33	1.89	0.11	Cytoplasmic dynein 1 light intermediate chain 1
30749633	-	6	1.78	0.22	Tyrosine-protein phosphatase non receptor type 1
38570062	-	5.81	0.61	0.61	UPF0363 protein C7 or f20
24307879	-	5.35	1.47	0.21	Coatomer subunit beta
	-	7.25	1.58	0.42	UTP-glucose-1-phosphate uridylyltransferase
					Overlap from PM to ER
	all[c]	*Top*[c]	*ND*[c]	*all*[c]	
4506773	-	5.26	-	-	**Protein S100-A9**
18655500	-	4.24	-	-	tr|Q6GMX0|Q6GMX Human Putative uncharacterized protein
12054072	-	3.03	-	-	Ig gamma-1 chain C region
5454024	-	2.99	-	0.37	Ribonuclease P protein subunit p30
4826659	0.36	2.89	-	0.36	F-actin-capping protein subunit beta
22726186	-	2.65	-	0.38	Proteosome assembly chaperone 2
13876386	-	2.51	-	0.73	Epiplakin
					Overlap from ER to PM
	all[c]	*ND*[c]	*Top*[c]	*all*[c]	
4506645	-	-	11.43	-	60S ribosomal protein L38
51036603	-	-	4.17	-	Guanine nucleotide-binding protein G(I)/G(S)/G(O) gamma-12
4506761	-	-	4.12	-	Protein S100-A10
4507129	-	-	4	2	Small nuclear ribonucleoprotein E
5454090	-	-	3.75	2	Translocon- associated protein subunit delta
6005860	-	-	3.2	2	60S ribosomal protein L35
7661728	-	-	3.2	-	Mitogen-activated protein binding protein interacting protein
					Overlap from ER to MT
	all[c]	*all*[c]	*Top*[c]	*ND*[c]	
4506645	-	-	11.43	-	60S ribosomal protein L38
10190712	-	0.96	8.65	-	Protein S100-A14
150010589	-	0.8	7.2	-	Interferon-induced transmembrane protein 1
17933772	-	2	4.8	-	Protein S100-A16
51036603	-	-	4.17	-	Guanine nucleotide-binding protein G(I)/G(S)/G(O) gamma-12
4506761	-	-	4.12	-	Protein S100-A10
3462883	-	1.32	3.51	-	Vesicle transport protein SEC20
					Overlap from MT to ER
	all[c]	*all*[c]	*ND*[c]	*Top*[c]	
1483131	-	0.34	-	12.24	**Nucleophosmin**
8922331	-	-	-	11.49	Protein mago nashi homologue 2
34201	-	-	-	10.91	60S ribosomal protein L35a
399758	-	-	-	9.52	Heterogenous nuclear ribonucleoprotein A3
7706425	-	-	-	9.38	U6 snRNA-associated Sm-like protein LSm8
11037094	-	-	-	8.7	tr|Q9HC85|Q9HC85 Human Metastasis realted protein
1232077	-	1.44	-	8.19	DNA replication licensing factor MCM2

a Proteins where the name is shown in bold correspond to proteins which exemplify general characteristics of the data that are noted in the text.
b Normalized abundances were calculated from the Spectral Abundance Factor using GeneSpring, that is, the abundances have been normalized using a correction for the differing number of amino acids in the proteins. For all proteins, the normalized abundances ranged from 0.018 to 22.25. A dash indicates the protein was not detected.
c Selection criteria. A filter to select non detected proteins was applied to a chosen fraction (ND). In an adjacent fraction in the sucrose gradient, the proteins were sorted according to abundance and the seven most abundant proteins (top) are shown.

Table 2. Test for overlap of proteins between sucrose gradient fractions

endoplasmic reticulum, mitochondria), while the databases typically use (cytoplasm/ cytosol, endoplasmic reticulum, mitochondrion, Golgi apparatus). Overall, relative to the UniProtKB subcellular locations, 271 proteins had no annotations, 1388 had annotations at the top level and 525 had annotations at the lower level. For the 481 (22.0%) proteins in the initial data set that were observed in only a single fraction, we compared their locations with previous experimental information about subcellular location in the UniProtKB database. Figure 6 summarizes the proportion of these "unique" proteins which were previously assigned to various subcellular locations. This data provides an overview of the enrichment of the four fractions with cytosolic, plasma membrane, endoplasmic reticulum and mitochondrial proteins respectively. First, all four fractions show a substantial proportion of proteins either for which there is no previous annotation of subcellular location, or for which the previous annotation is only nucleus or extracellular region (from 5 (19%) of proteins in ER to 83 (40%) of proteins in MT). These annotations are compatible with the enrichment of the fractions with their various types of proteins and the present results constitute new annotation information for these proteins. Fraction CT shows three other major slices: (1) proteins which are fully compatible with cytosolic proteins, (2) proteins which have previously been assigned to cytoplasm, but also to other subcellular locations, and (3) proteins which have been previously assigned to other subcellular locations, but not to cytoplasm or cytosol. There is some ambiguity in the second and third groups since cytosol is not distinguished in many experimental strategies and the assigned locations are daughters of cytoplasm (but not of cytosol) in the GO ontology. Overall for the 127 proteins observed only in fraction CT, 119 (93.7%) have annotations that are compatible with enrichment of this fraction with cytosolic proteins. Only 8 proteins (6.3%) appear to be discrepancies that have other, incompatible locations. Of the 119 compatible proteins, 16 proteins have previous annotations that deviate from observation uniquely in fraction CT. For the other sucrose gradient fractions the cytoplasm/cytosol distinction also leads to some ambiguity, but overall the number/proportion of proteins compatible with enrichment of fraction PM (plasma membrane), fraction ER (endoplasmic reticulum) and fraction MT (mitochondrion) with the respective protein types are 94 (78.3%), 18 (67.0%), and 184 (88.9%) respectively. Because there is some inconsistency between the different subcellular location annotation sources , these numbers vary somewhat if the UniProtKB subcellular components or the Subcellular Location database are used, but do not change the overall conclusion. Within the limitations of such comparisons, we conclude that the previous annotations are largely consistent with enrichment of the fractions with the expected protein types. Apparent experimental/ database annotation discrepancies for all 2184 proteins are considered in more detail below. Is the apparent multiplicity of protein subcellular locations observed in our experiments captured in current database annotations? To address this question, we used the set of 163 proteins in the initial data set that showed bimodal, nonadjacent distributions over the sucrose gradient fractions (includes proteins observed only in combinations of non-adjacent fractions CT-ER, CT-MT, PM-MT, CT-PM-MT, and CT-ER-MT, i.e. proteins that clearly have multiple locations) and which also had at least 8 spectral counts. The latter condition ensures that the classification of these proteins as bimodal is not unduly influenced by the dynamic range limitations of MS/MS spectral counting. This set of proteins was compared with (merged) subcellular location annotations from the UniProtKB and LOCATE Subcellular Location databases. Figure 7 shows the distribution over the bimodal combinations of fractions and the annotations of subcellular location for all 163 proteins. As seen above with the proteins identified in only a single

fraction, 59 (36.2%) of the bimodal proteins only had annotation at the level (nucleus, extracellular region, no annotation). Furthermore, only 22 (13.5%) of the proteins show multiple locations at the annotation level (cytoplasm/cytosol, plasma membrane, endoplasmic reticulum, Golgi apparatus, mitochondrion). In general these results are consistent with the conclusion that current database annotations of subcellular location are sparse and skewed toward single locations for proteins.

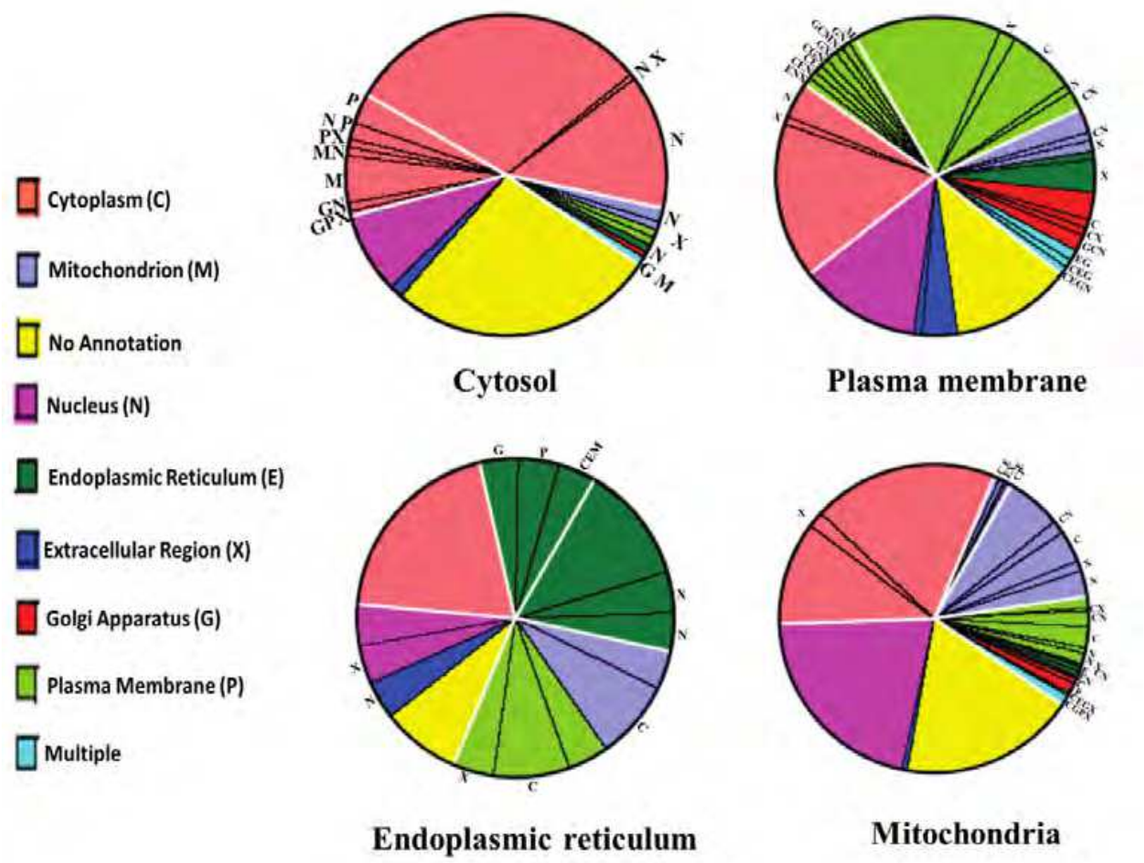


Fig. 6. Distribution of current subcellular location annotations in the UniProtKB database over the proteins observed solely in a single sucrose gradient fraction in the initial data set. The annotations are color coded (legend) according to the GO classification levels compatible with our experimental strategy (upper: extracellular region, plasma membrane, cytoplasm, nucleus; lower: cytosol/cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion). A small number of proteins had multiple lower level annotations and are shown in the region color coded as multiple. Proteins that had multiple annotations that included other locations different from the color code are indicated by the radial letters. The heavy white lines delineate slice regions that have different compatibility with the experimental data.

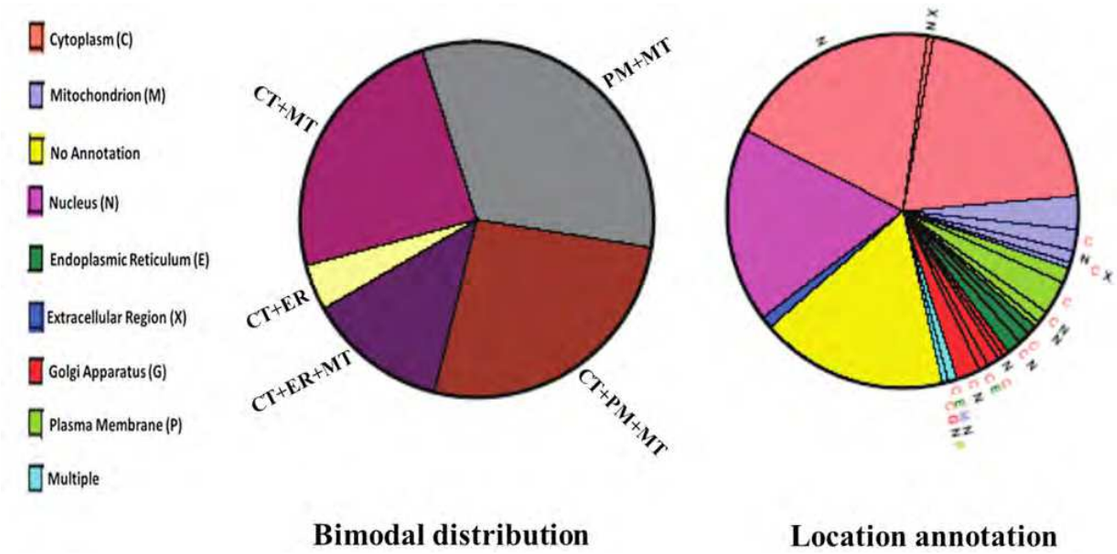


Fig. 7. Comparison of the present data on subcellular location of bimodal proteins with (merged) subcellular location annotations in the UniProtKB subcellular location comments, UniProtKB subcellular component GO terms and LOCATE Subcellular Location database. (left) Distribution of the 299 bimodal proteins in the initial data set over different combinations of sucrose gradient fractions. The indicated combinations of fractions can only arise for proteins with at least two different subcellular locations. (right) Summary of the (merged) subcellular location annotations for all 299 bimodal proteins. The annotations are color coded (legend) according to the GO classification levels compatible with our experimental strategy (upper: extracellular region, plasma membrane, cytoplasm, nucleus; lower: cytosol/cytoplasm, endoplasmic reticulum, Golgi apparatus, mitochondrion). A small number of proteins had multiple lower level annotations and are shown in the region color coded as multiple. Proteins that had multiple annotations that included other locations different from the color code are indicated by the radial letters. Slices that have color coded radial lettering are those corresponding to proteins whose annotations indicate multiple subcellular locations within the GO classification levels of our experimental design.

Over all of the 2184 proteins, the annotations at the subcellular level in the examined databases tend to be to single locations. Given that many previous proteomics studies were biased against detection of proteins in multiple locations (e.g., studies of purified organelles) and that annotations at sub-cytoplasmic levels are clearly still very sparse, we consider that the previously available annotations of experimental data are not inconsistent with the proposal that many, probably a sizable majority, of the proteins have multiple subcellular locations. Using the initial data set of (protein, fraction) pairs, there were a relatively small number of discrepancies between our data and previous annotations of subcellular location in the two databases. Of the 1441 proteins identified in fraction ER, there were a total of 33 proteins previously annotated to endoplasmic reticulum that we did not observe in fraction ER. Similarly for fractions PM (1611 proteins) and MT (1610 proteins), there were a total of 58 and 29 proteins previously annotated to plasma membrane and mitochondrion respectively that we did not observe in the corresponding gradient fraction.
Inconsistencies in the databases might contribute to the apparent discrepancies. For the 2184 proteins identified here, Figure 8 shows the status of annotations of plasma membrane (443

proteins), mitochondrion (168) and endoplasmic reticulum (243) proteins. There is rather little concordance between the annotation sets, which presumably must reflect the inclusion of very different experimental data sets.

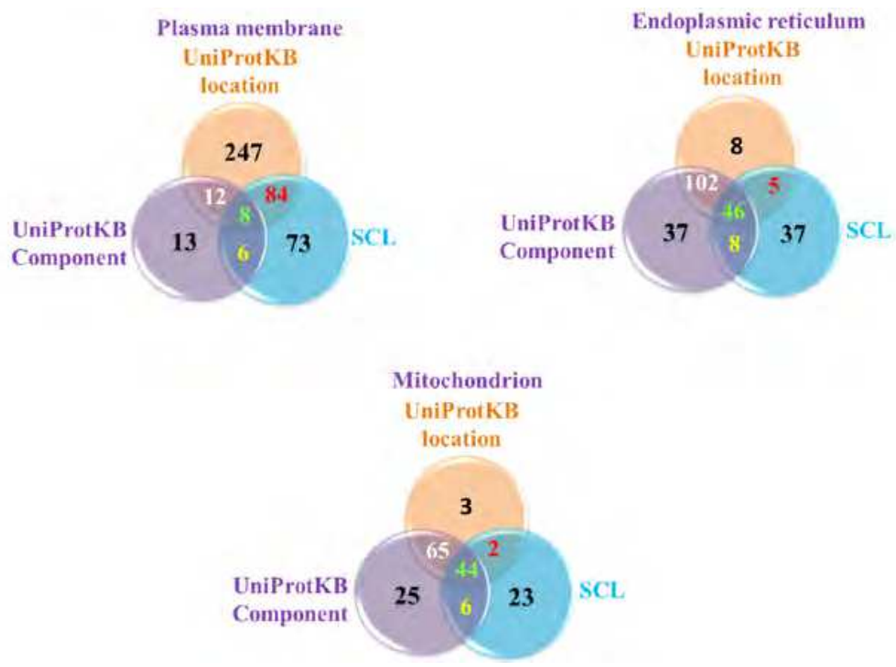


Fig. 8. Venn diagrams comparing the status of annotations of subcellular locations in the UniProtKB subcellular location components, UniProtKB cellular component GO terms and LOCATE Subcellular Location database for proteins observed in this work. 443 proteins annotated as plasma membrane. 168 proteins annotated as mitochondrion. 243 proteins annotated as endoplasmic reticulum.

Only 8 of the 443 proteins with annotations of plasma membrane were so annotated in all three data sets! For the proteins annotated to plasma membrane, endoplasmic reticulum and mitochondrion that we did not observe in the corresponding gradient fractions, our data would suggest different primary locations for these 120 proteins, but does not exclude their presence in the annotated subcellular locations as secondary locations which could not be detected at our sensitivity limits. We believe that some occurrences of apparent discrepancies are almost inevitable for three reasons. First, there is still very little information about whether subcellular distributions of proteins are the same in different cell types or under different cellular conditions. Second, many experiments do not distinguish between different isoforms of the same protein, which may have different subcellular distributions. Indeed, the present data set includes these proteins, which in part show different distributions over subcellular locations for isoforms of the same protein. This data will be analyzed in a separate paper. Third, the databases attempt to aggregate data from experimental strategies with very different sensitivity, selectivity, dynamic range, and coverage of proteins. Targeted searches for individual proteins in purified subcellular fractions with antibody methods probably have the highest sensitivity for detecting trace amounts of proteins in any specified location, even if the trace is a tiny proportion of the total protein abundance. Conversely, some high throughput methods may have limited

resolution for some subcellular locations, for example, distinguishing cytosol from cytoplasm, and may have insufficient sensitivity and dynamic range to detect trace amounts of proteins in specific locations. Aggregating subcellular location information from many cell types and conditions obtained with very different experimental strategies, many of which do not distinguish protein isoforms, then becomes a very tricky task which seems likely to produce some discrepancies with any specific experimental method/data set. Although only a few of the fractions from the sucrose density gradient have been analyzed, the normal data set provides clear evidence that a minimum of 543 of the 2184 proteins (24.9%) show multiple locations. The minimum estimate is based on those proteins that are either present in all fractions or show bimodal distributions with abundance peaks in nonadjacent fractions of the sucrose gradient (Figure 3B, C). For the 321 proteins (14.7%) that were found only in adjacent fractions of the gradient (i.e., CT-PM, PM-ER, and ER-MT), the present experiments are insufficient to exclude that this might be due to the presence of a single organelle that occupies an intermediate position between the two fractions. On the other hand, we intentionally spaced the analyzed fractions widely in the sucrose gradient and for the 476 proteins (21.8%) that were found in three adjacent fractions (i.e., CT-PM-ER, or PM-ER-MT), it is improbable that these proteins have single subcellular locations. Especially since other proteins demonstrated lack of overlap (e.g., proteins in fractions CT-ER or PM-MT) and lack of tailing in the sucrose gradient (Table 2). Furthermore, in most cases the relative abundances for the proteins observed in three adjacent fractions were substantial and did not correspond to trace proportions. Thus, the normal data set provides evidence indicating that 38.6% of the proteins may have unique locations, 24.9% certainly have multiple locations, 21.8% most likely have multiple locations and 14.7% may have either unique or multiple locations. We have used the observed set of proteins to examine possible connections between subcellular location and function as related to breast cancer. Many of the proteins observed in our experiments have previously been annotated with functional information.

2.1.6 Breast cancer related proteins

Biological process annotations for 1673 proteins, molecular function annotations for 1980 proteins, Reactome Pathway annotations for 176 proteins and posttranslational modification annotations for 1653 proteins were available in the UniProtKB database. We used the BioBase Biological Databases, BIOBASE Knowledge Library (BKL) and ExPlainTM 2.3 platform to identify 94 proteins in our data set that are known or suspected to be implicated in breast cancer via disease molecular mechanism, diagnostic marker and therapeutic target association. These proteins were examined for common Gene Ontology (http://www.geneontology.org) biological process and molecular function terms and for common Reactome Pathway (http://www.reactome.org) terms, which were then used as lures to obtain the set of proteins identified in this study that share the same terms. A majority of the proteins implicated in breast cancer were related to five high level cellular processes that involved a subset of 519 proteins observed in our experiments: apoptosis (68 proteins), cell growth (127), signaling (131), cell interaction (62), and protein processing (230). 93 proteins were involved in more than one of the five processes. Figure 9 shows how the proteins associated with each cellular process are distributed over the subcellular locations using the initial data set. The striking features are that each process is distributed over all four locations, as might be anticipated for regulated processes, and that for all of the cellular processes there is an appreciable majority of proteins with 3-4 subcellular locations

(ranging from 54.8% for cell interaction to 66.5% for protein processing). Furthermore, the latter characteristic was most pronounced for the 93 proteins that were involved in more than one of the high level cellular processes (68.8%)

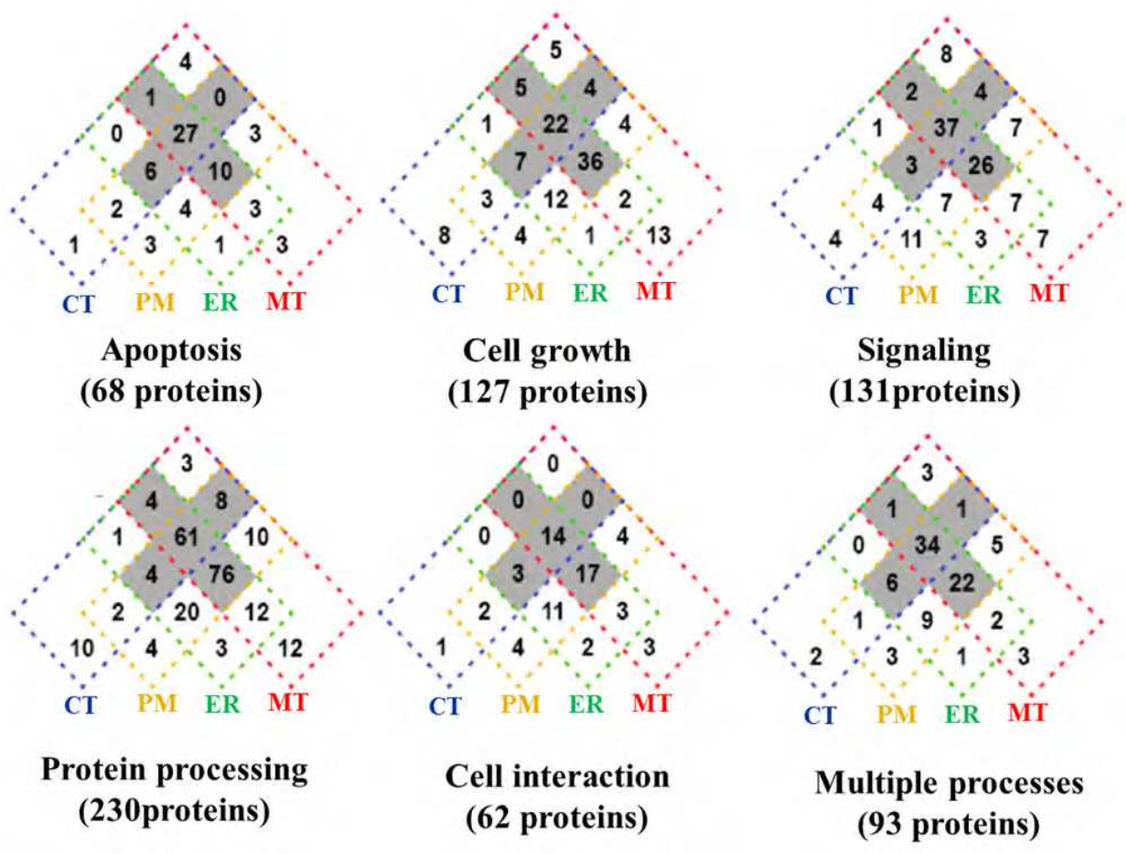


Fig. 9. Four-way Venn diagrams summarizing the distribution of the breast-cancer-related set of 519 proteins over the subcellular locations for the cellular processes: signaling (131 proteins), cell growth (127), protein processing (230), apoptosis (68 proteins), and cell interaction (62), as well as for proteins involved in more than one of these cellular processes (93). The shaded regions of the diagrams correspond to proteins with 3 or 4 locations.

3. Conclusion

Large scale and quantitative proteomics analysis of subcellular organelles revealed 268 nuclear proteins and 22 extracellular region proteins that were found in various sucrose gradient fractions, but which had previously only been annotated experimentally to the nucleus and extracellular region respectively. Another 271 proteins that we detected had no prior annotation at either the upper level (cytoplasm, plasma membrane) or lower level (cytosol, endoplasmic reticulum, mitochondria) of our experimental strategy. The present experiments were not designed to obtain specific annotations at the lower level, e.g. to mitochondrion. Hence, observation of a protein in a Fraction MT that is enriched in mitochondrial proteins should presently only be taken as an indication and not as proof of its presence in mitochondria. Nonetheless, the present experiments gave several hundred

new location annotations at the level (plasma membrane, cytoplasm). There are several ways the limits on MS detection sensitivity may influence the number of locations in which the proteins were observed. In particular, for the highest abundance proteins, the sensitivity and dynamic range of the MS spectral counting methods are such that trace amounts as small as about 0.2% of a protein in a secondary location could be detected. As shown above for the normal data set, trace amounts of abundant proteins in secondary locations do not strongly influence estimates of the proportion of proteins with multiple subcellular locations. Conversely, the proportion of a protein which must be present in a secondary location to be detectable increases as the overall abundance of the proteins decreases, for example, for the lowest abundance proteins, only the highest abundance, primary location falls within the detection limits of the MS methods. Furthermore, for lower abundance proteins or for trace proportions of proteins in specific fractions, the sampling constraints on spectral counting that result from MS/MS sequencing of only the more abundant peptides means that only one peptide may be counted in some fractions. For example, there were 847 (38.8%) proteins classified as "unique" (observed in a single fraction) in the normal data set, but only 481 (22.0%) in the initial data set. This difference corresponds to proteins in various gradient fractions that were only counted with a single peptide and 1 or 2 spectral counts. This means that estimations of multiple locations based on the normal data set are very conservative and certainly underestimate, probably strongly, the proportion of proteins with multiple subcellular locations. Given that estimates based on the normal data set provide evidence for multiple locations of at least 46.7% of the observed proteins, we conclude that a substantial majority of the proteins observed have multiple subcellular locations. Given that only 22% of proteins were seen solely in a single fraction in the initial data set, perhaps as much as 75% of the proteins have multiple locations. We noted above that 120 proteins had annotations to subcellular locations that we did not observe in the corresponding sucrose gradient fractions (33 to endoplasmic reticulum, 58 to plasma membrane and 29 to mitochondrion). We suggested that these discrepancies were not inconsistent with our data if the annotations corresponded to secondary locations. On the basis of the observed spectral counts, there are 39 of these proteins for which our data suggest that the previous annotations correspond to proteins with functional significance in a secondary location, but that >80% of the protein is in a different primary location. This kind of analysis can be extended to many other proteins where the functional activity and the measured mole fractions indicate functional roles at secondary locations. Indeed, some of the proteins that we detected at trace amounts (<3%) in secondary locations already have known functions at those locations. The present experiments thus indicate numerous proteins with primary locations which probably differ from current function/location annotations and for which confirmation of the primary location (and potentially of other functional activities) might be profitably sought. The present experiments suggest 1383 (protein, location, function) data points for 519 proteins involved in five major cellular functional processes for which investigation of functional roles might further elucidate mechanisms involved in breast cancer. This is a very promising situation for experiments aimed at investigating dynamic changes in the spatio/temporal location/form of proteins in breast cancer cells, their potential roles in regulation and their potential importance in breast cancer disease. Finally, in summary, we have found evidence that strongly suggests a majority of the detected proteins have multiple subcellular locations in the breast cancer model MCF-7 cells, that even with a fairly simple experiment a wealth of new annotation data can be obtained, that available evidence suggests that for many proteins distribution

over multiple subcellular locations can be important to their functional roles, and that large numbers of (protein, location) pairs deserving of further investigation of functional/regulatory roles can be delineated. We are still very far from having good static descriptions of the spatial distributions of cellular proteins, let alone dynamic information on relationships between spatio/temporal distribution and function. However, high-throughput proteomics in combination with other experimental methods seems to offer ways forward.

4. Acknowledgment

We would like to thank Professor Samia M. A. Al-Amoudi (King Abdulaziz University, Jeddah, SA) for generous help and discussions and to the Wellcome Trust, UK, King Faisal Foundation, Riyadh, SA, Sheikh Mohammed Hussien Al-Amoudi Centre Scientific Chair for Breast Cancer Researches, Jeddah, SA, Susan G. Komen for the Cure Breast Cancer Foundation, USA, King Faisal Specialist Hospital and Research Center, Riyadh, SA for their continuous support.
Figs, tables and part of the text reprinted and adapted with permission from Qattan et al (2010), J Proteom Res 9, 459-508. Copyright (2011) American Chemical Society.

5. References

http://www.ncbi.nlm.nih.gov
Au, C. E., A. W. Bell, A. Gilchrist, J. Hiding, T. Nilsson & J. J. Bergeron (2007) Organellar proteomics to create the cell map. *Curr Opin Cell Biol,* 19, 376-85.
Cheyne, W. W. (1894) An Address on the Treatment of Cancer of the Breast: Delivered before the Harveian Society of London. *Br Med J,* 1, 289-91.
Dunkley, T. P., R. Watson, J. L. Griffin, P. Dupree & K. S. Lilley (2004) Localization of organelle proteins by isotope tagging (LOPIT). *Mol Cell Proteomics,* 3, 1128-34.
Foster, L. J., C. L. de Hoog, Y. Zhang, X. Xie, V. K. Mootha & M. Mann (2006) A mammalian organelle map by protein correlation profiling. *Cell,* 125, 187-99.
Godovac-Zimmermann, J., O. Kleiner, L. R. Brown & A. K. Drukier (2005) Perspectives in spicing up proteomics with splicing. *Proteomics,* 5, 699-709.
Gygi, S. P., Y. Rochon, B. R. Franza & R. Aebersold (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol,* 19, 1720-30.
Hanahan, D. & R. A. Weinberg (2000) The hallmarks of cancer. *Cell,* 100, 57-70.
Hanahan, D. & R. A. Weinberg (2011) Hallmarks of cancer: the next generation. *Cell,* 144, 646-74.
Imyanitov, E. & K. Hanson (2004) Mechanisms of breast cancer. *Drug Discovery Today: Disease Mechanisms,* 1, 235-245.
Ince, T. A. & R. A. Weinberg (2002) Functional genomics and the breast cancer problem. *Cancer Cell,* 1, 15-7.
Lacroix, M. & G. Leclercq (2004) Relevance of breast cancer cell lines as models for breast tumours: an update. *Breast Cancer Res Treat,* 83, 249-89.
Mukherjee, S. 2011. *The Emperor of All Maladies: A Biography of Cancer*. UK: Fourth Estate Ltd.

Qattan, A. T., C. Mulvey, M. Crawford, D. A. Natale & J. Godovac-Zimmermann (2010) Quantitative organelle proteomics of MCF-7 breast cancer cells reveals multiple subcellular locations for proteins in cellular functional processes. *J Proteome Res,* 9, 495-508.

Roberts, G. C. & C. W. Smith (2002) Alternative splicing: combinatorial output from the genome. *Curr Opin Chem Biol,* 6, 375-83.

Rogers, L. D. & L. J. Foster (2007) The dynamic phagosomal proteome and the contribution of the endoplasmic reticulum. *Proc Natl Acad Sci U S A,* 104, 18520-5.

Sebastian, T. & P. F. Johnson (2006) Stop and go: anti-proliferative and mitogenic functions of the transcription factor C/EBPbeta. *Cell Cycle,* 5, 953-7.

Simpson, J. C. & R. Pepperkok (2006) The subcellular localization of the mammalian proteome comes a fraction closer. *Genome Biol,* 7, 222.

Sorlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning & A. L. Borresen-Dale (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A,* 98, 10869-74.

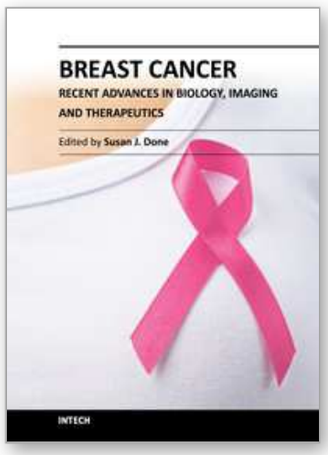
Soule, H. D., J. Vazguez, A. Long, S. Albert & M. Brennan (1973) A human cell line from a pleural effusion derived from a breast carcinoma. *J Natl Cancer Inst,* 51, 1409-16.

Usaite, R., J. Wohlschlegel, J. D. Venable, S. K. Park, J. Nielsen, L. Olsson & J. R. Yates Iii (2008) Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. *J Proteome Res,* 7, 266-75.

van de Vijver, M. J., Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend & R. Bernards (2002) A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med,* 347, 1999-2009.

Xu, P., M. Crawford, M. Way, J. Godovac-Zimmermann, A. W. Segal & M. Radulovic (2009) Subproteome analysis of the neutrophil cytoskeleton. *Proteomics,* 9, 2037-49.

Yates, J. R., 3rd, A. Gilchrist, K. E. Howell & J. J. Bergeron (2005) Proteomics of organelles and large cellular structures. *Nat Rev Mol Cell Biol,* 6, 702-14.



Breast Cancer - Recent Advances in Biology, Imaging and Therapeutics

Edited by Dr. Susan Done

ISBN 978-953-307-730-7
Hard cover, 428 pages
Publisher InTech
Published online 14, December, 2011
Published in print edition December, 2011

In recent years it has become clear that breast cancer is not a single disease but rather that the term encompasses a number of molecularly distinct tumors arising from the epithelial cells of the breast. There is an urgent need to better understand these distinct subtypes and develop tailored diagnostic approaches and treatments appropriate to each. This book considers breast cancer from many novel and exciting perspectives. New insights into the basic biology of breast cancer are discussed together with high throughput approaches to molecular profiling. Innovative strategies for diagnosis and imaging are presented as well as emerging perspectives on breast cancer treatment. Each of the topics in this volume is addressed by respected experts in their fields and it is hoped that readers will be stimulated and challenged by the contents.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen